
神经网络的弱点-对抗攻击及其防御

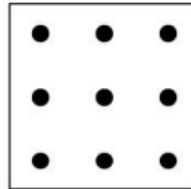
唐呈俊 t@tcj.ac.cn *Guilin University of Electronic Technology*

August 17, 2020

目录

- ① 攻击模式分类
- ② 基础攻击对抗样本生成方式
- ③ 对抗攻击防御
- ④ 基于GAN的对抗样本生成方式
- ⑤ 其它对抗攻击和防御方式及思考
- ⑥ 实体攻击
- ⑦ 谢谢

想想看?



- 2.一个房间的天花板上吊着两根绳子. 两根绳子相隔很远, 无法同时抓住. 房间里只有一把椅子、一盒火柴和一把钳子, 怎么才能把两根绳子系在一起呢.
- 3.数学证明题中, 三角形ABC和三个点ABC有什么区别呢?

Figure. 想想看怎么才能用一笔画出四个线段将这九个点全部连接起来呢?

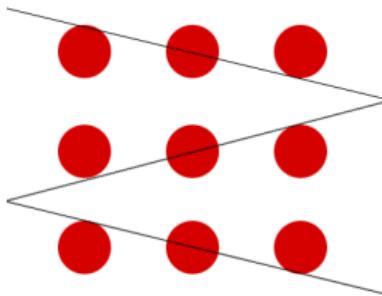
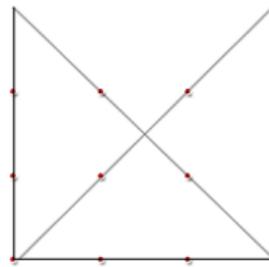


Figure. 打破条框即可

2. 把钳子作为重物系在一根绳子上，使绳子形成单摆运动，当两根绳子靠得很近时，抓住另外一根绳子，从而把两根绳子系起来。

3. 非退化条件(nondegenerate condition)是使几何命题不失一般性的条件。

输出若干附加条件，并断言在这些附加条件下命题的结论成立。

引子

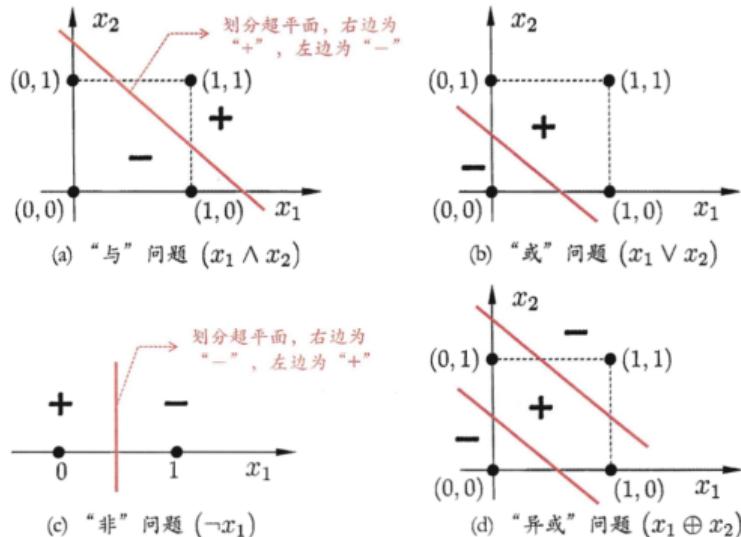


Figure. 线性可分问题

神经网络最早可以追溯至Rosenblatt感知器, 它可以看作是只有一层而没有隐藏层的神经网络, 它只可以处理线性可分的问题. 后来随着反向传播算法(BP算法)的发明, 多层感知机进入了我们的视野, 其强大的拟合能力使得其可以拟合绝大部分的函数.

攻击模式分类

模型透明度

透明项	模型参数	计算图	数据集	模型输出	预测标签
白盒攻击	有	有	有	有	有
黑盒攻击	无	无	无	无	有

当然,还存在许多透明度状况在黑盒攻击和白盒攻击之间的情况.

攻击目的

无目标攻击

对于一张图片，生成一个对抗样本，使得标注系统在其上的标注与原标注无关，即只要攻击成功就好，对抗样本的最终属于哪一类不做限制。

有目标攻击

对于一张图片和一个目标标注句子，生成一个对抗样本，使得标注系统在其上的标注与目标标注完全一致，即不仅要求攻击成功，还要求生成的对抗样本属于特定的类。

攻击方法

数字攻击

能够直接提交模型输入, 生成的扰动可以非常精确地输入到模型.

实体攻击

无法直接向模型提交输入, 只能通过其它间接的方式对模型输入进行扰动.

基础攻击对抗样本生成方式

基于梯度的方法FGSM

FGSM方法利用了模型的梯度信息, 故单纯的FGSM仅能在白盒攻击中使用.

$$x + \varepsilon \operatorname{sgn}(\nabla_x L(\theta, x, y))$$

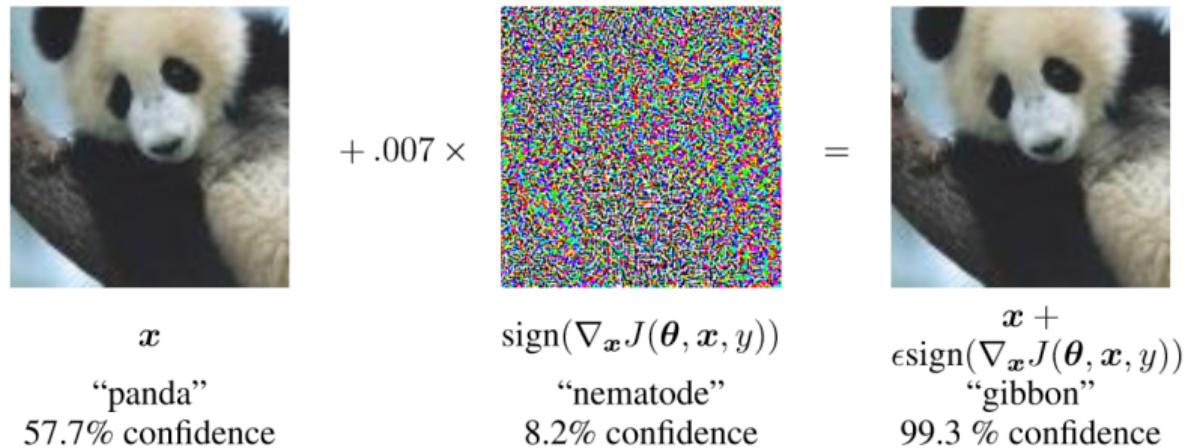


Figure. Fast gradient sign method

迭代+梯度 PDG (projected gradient descent)

$$x^{t+1} = \Pi_{x+S} \left(x^t + \alpha \operatorname{sgn}(\nabla_x L(\theta, x, y)) \right)$$

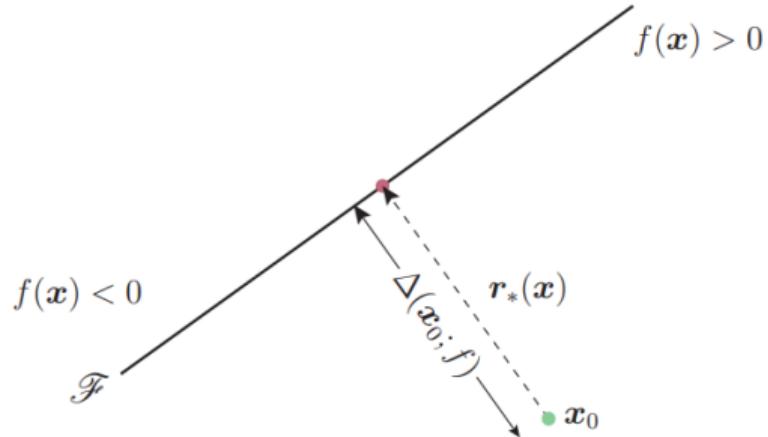


Figure. 范数约束(分离超平面) DeepFool

对抗攻击防御

对抗训练

在模型训练过程中，训练样本不再只是原始样本，而是原始样本加上对抗样本。对抗训练本身可以被表述为一个Min-Max优化。

Min-Max优化

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\Delta x \in \Omega} L(x + \Delta x, y; \theta) \right]$$

对抗训练模型在干净数据上的准确率可以和原模型相近。

缺点

防御仍然局限于训练使用的对抗样本生成方法，对于黑盒攻击的防御性不高。(虽然可以通过使用多种模型的对抗样本输入来改进，但效果仍不够理想)

由于当前的许多对抗样本生成方法都是基于梯度去生成的，所以如果将模型的原始梯度信息隐藏或混淆起来，就可以达到抵御对抗样本攻击的效果。

防御性蒸馏

我们把数据结构信息和数据本身当作一个混合物，分布信息通过概率分布被分离出来。首先， T 值很大，相当于用很高的温度将关键的分布信息从原有的数据中分离，之后在同样的温度下用新模型融合蒸馏出来的数据分布，最后恢复温度，让两者充分融合。

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

缺点

本质仅仅只是替换了模型，使用其它方法仍然可以生成对抗样本。

防御性蒸馏

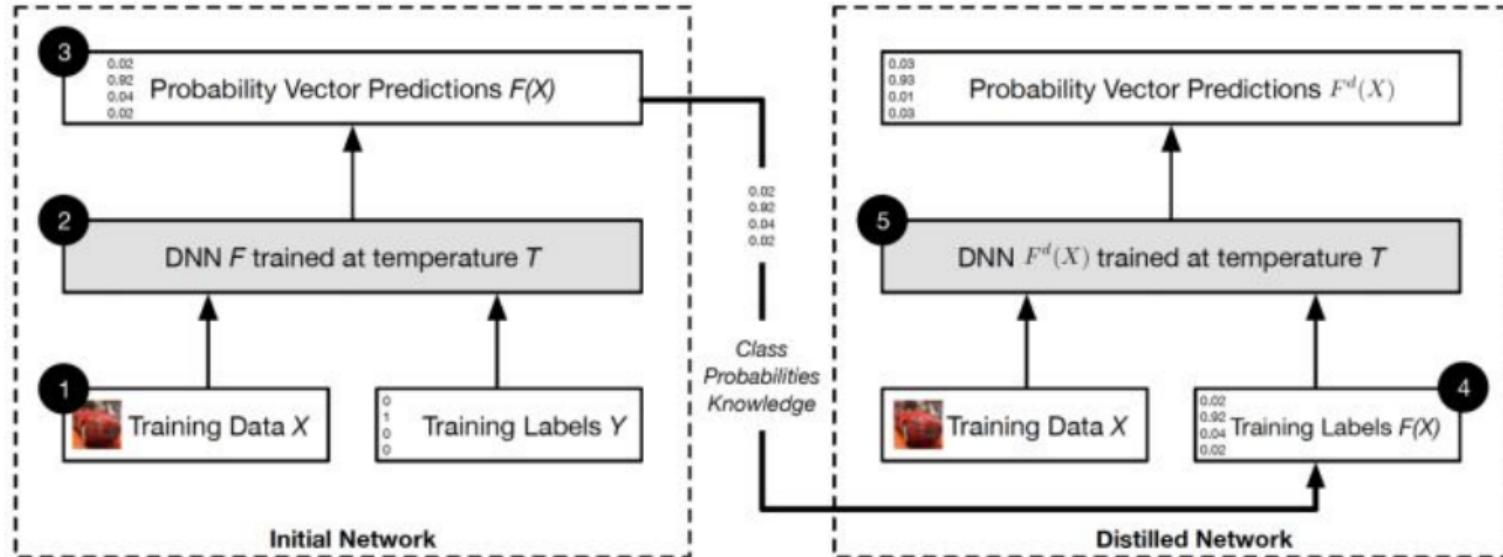


Figure. 防御性蒸馏策略

由于许多数字对抗攻击方法都是通过向输入中添加类似噪声的扰动来达到生成对抗样本, 因此降噪将会是低于对抗攻击的一个好方法.

对模型输入进行降噪

在数据输入模型之前就进行降噪操作, 理论上来说可以扰乱或去除对抗性噪声.

对输入进行降噪的缺点

传统的降噪方法可能会适得其反地将对抗性噪声反而放大, 因此提出[在模型中插入降噪模块](#)的方法.

传统的降噪方法分为以下几类:

空域滤波 领域平均滤波, 中值滤波

频域滤波 傅里叶变换滤波, 小波变换滤波

偏微分方程

变分法

形态学方法

降噪(模型)

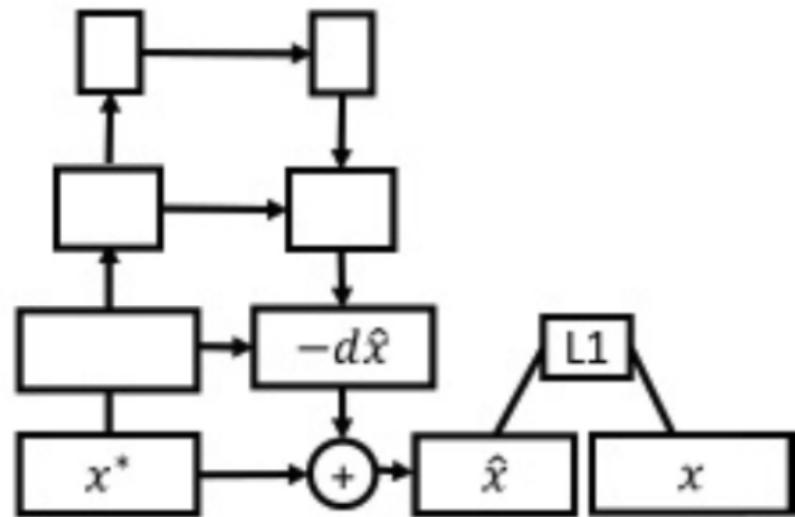
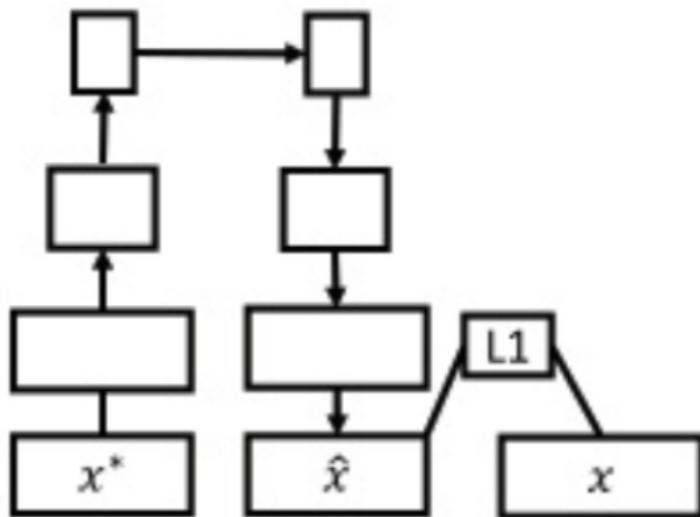


Figure. denoising Autoencoder (DAE) 和 Denoising Additive U-Net (DUNET)

降噪(模型)

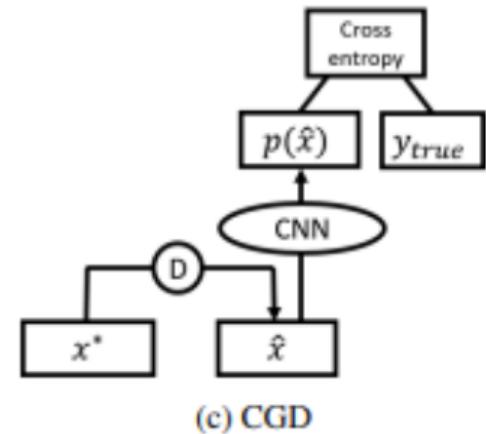
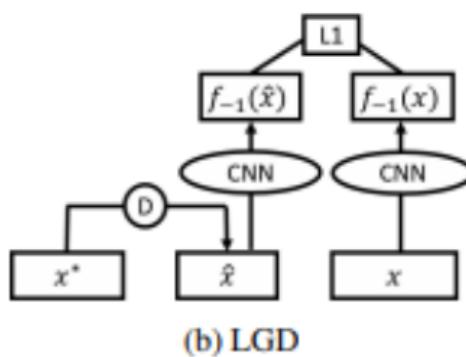
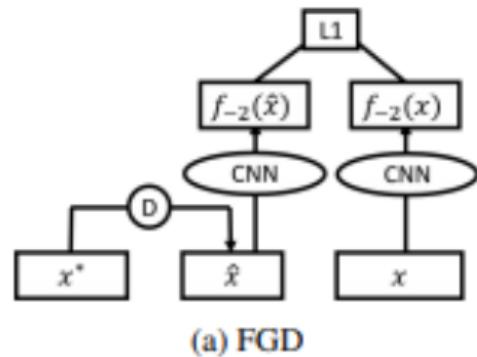


Figure. Denoising Additive U-Net (DUNET)

降噪(模型)

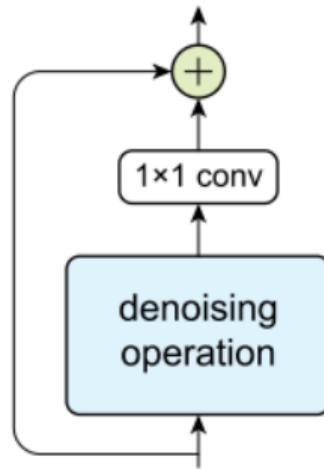


Figure. Denoising block

随机化

随机化方法和降噪类似, 同样可以在模型外和模型内完成.

(个人观点: 也许降噪也只是一种定向的随机化?)

主要随机化方法

预处理 随机对图像进行缩放旋转裁剪等变换和遮罩等.

Dropout Dropout不仅可以在训练中开启.

向原始模型引入随机层或者随机变量. 使模型具有一定随机性, 全面提高模型的鲁棒性, 使其对噪声的容忍度变高. 但会降低准确度.

基于**GAN**的对抗样本生成方式

使用对抗生成网络生成对抗样本

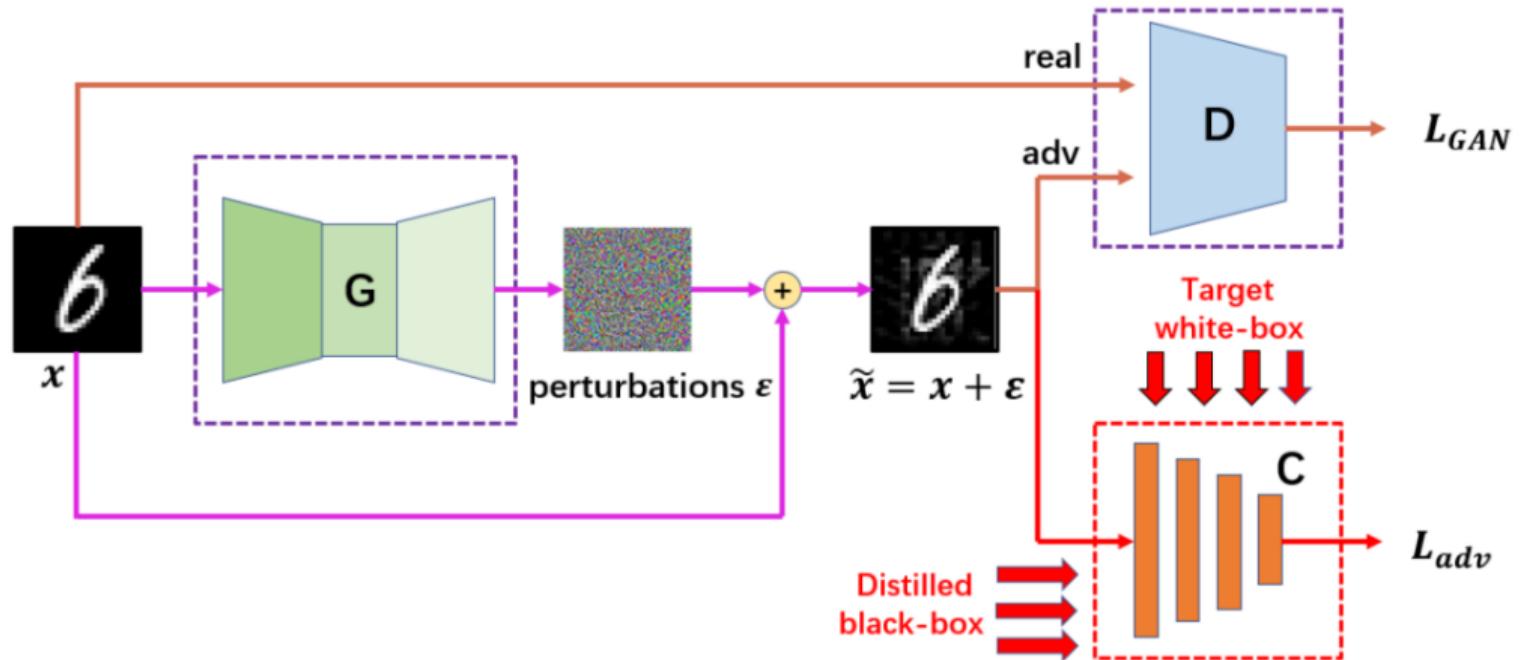


Figure. Generating Adversarial Examples with Adversarial Network

利用隐藏层信息生成对抗样本

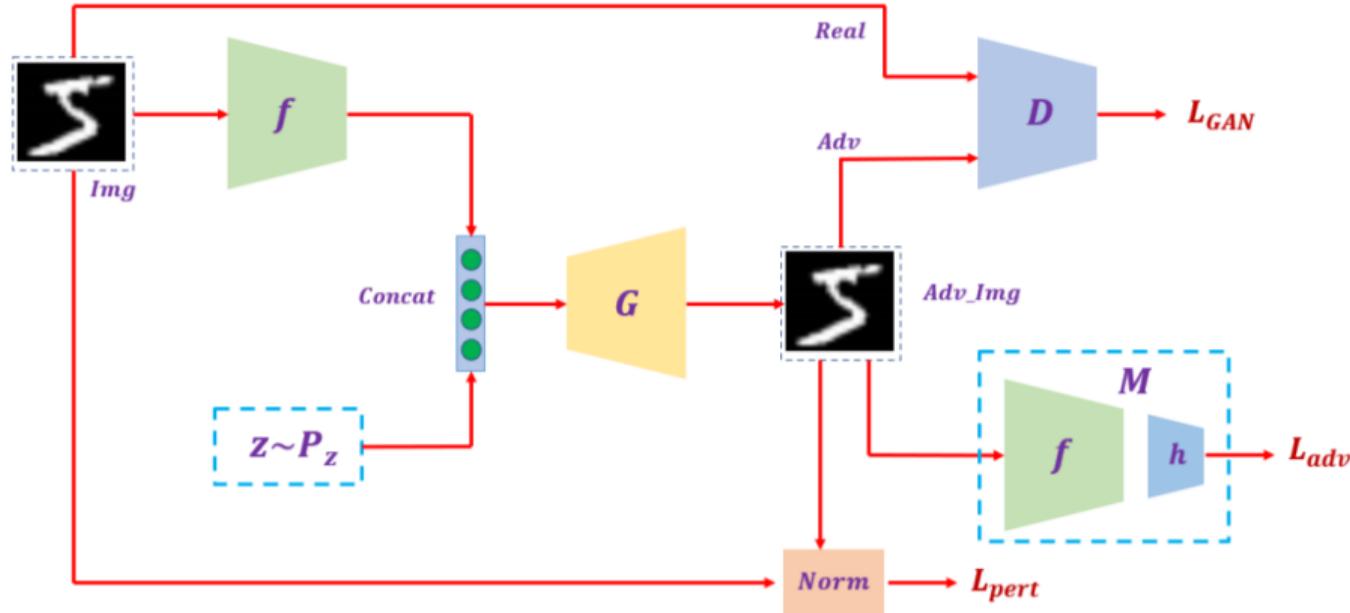


Figure. AdvGAN++: Harnessing latent layers for adversary generation

其它对抗攻击和防御方式及思考

使神经网络的功能重编程

通过向测试图片添加对抗扰动信息，当这些图片输入系统后，攻击模型将进行**重编程(reprogramming)**，进而抛弃本职任务而去进行其他任务（对抗任务）。

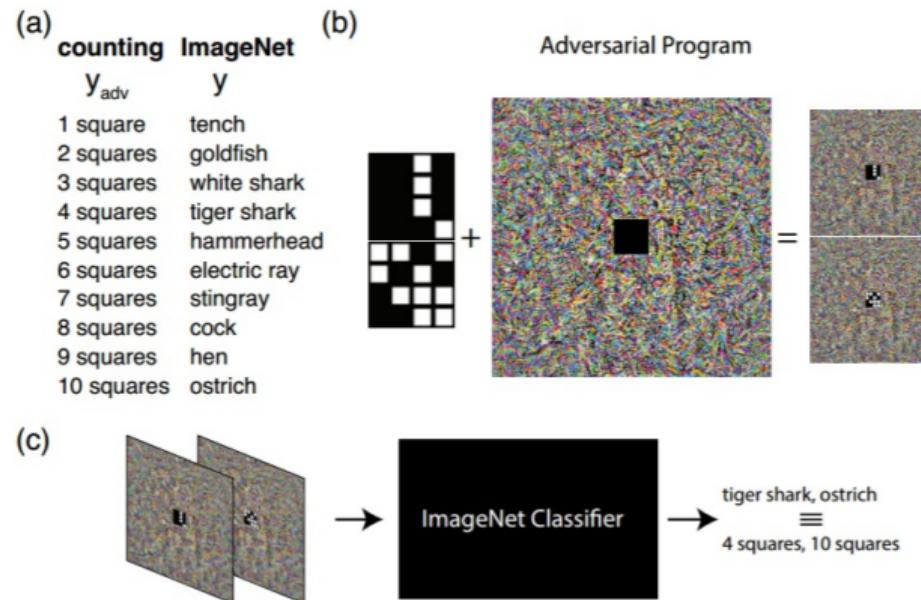


Figure. 将ImageNet分类器重编程为一个数方块的分类器

使用计算机图形学噪声进行攻击

计算机图形学中的噪声其实并不是通常所说的(白)噪声,而是一种生成纹理的方法,在过去计算机储存容量低下的时候,游戏中的纹理通常就是使用噪声算法生成的.



Figure. 魂斗罗中的纹理

使用计算机图形学噪声进行攻击

这些噪声算法都有一定的参数可以控制生成的纹理样式, 使用贝叶斯优化算法可以对这些参数进行优化, 从而找到对抗样本.

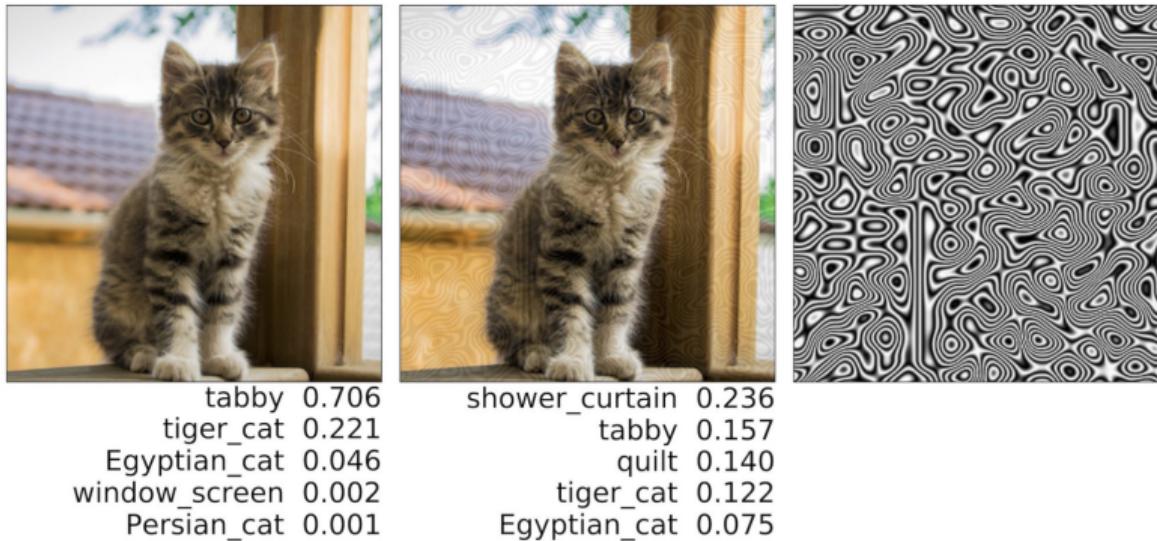


Figure. 使用Perlin噪声生成对抗样本

神经网络还不够强吗?

神经网络为什么会有弱点,难道是因为它还不够强吗?并不是,它的表达能力已经太强了.

继续思考

研究发现,在对抗训练过程中,即使将底层权重全部冻结,仅仅更新最顶层(softmax层)的参数同样能起到很好的效果.

可是这意味着什么?

实体攻击

贴纸攻击

在实体世界中, 难以施加对整个图像的扰动, 但可以使用贴纸的方式干扰神经网络的输出.

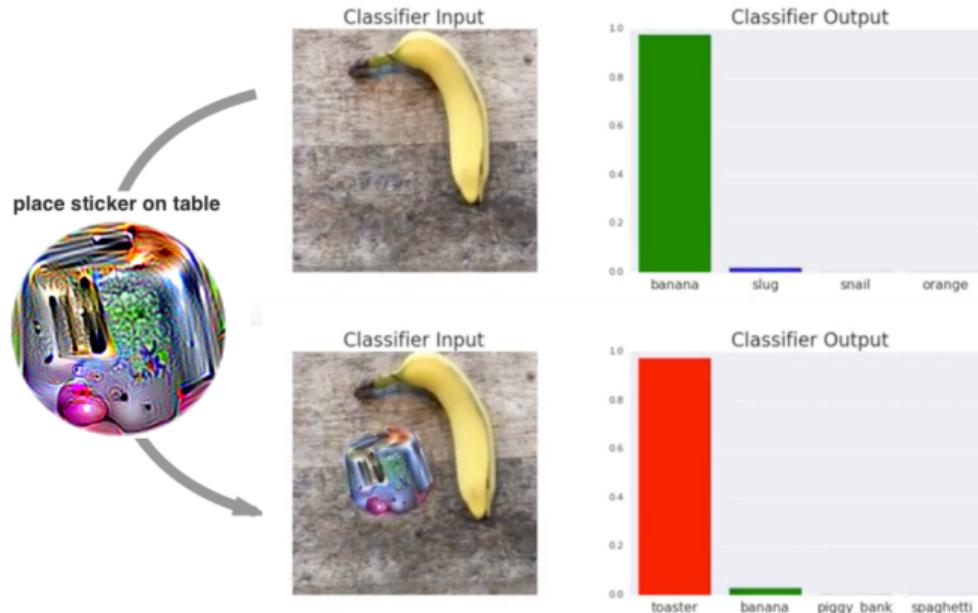


Figure. 使用对抗性贴纸进行攻击

对交通标识的实体攻击

自动驾驶过程中的交通标识识别非常重要, 而对交通标识的对抗攻击需要在不同角度和距离都能产生对抗攻击效果.



Figure. 对交通标识不同角度和距离的物理攻击

对交通标识的实体攻击



Figure. 或者对交通标识使用纯色贴纸

使用风格化方法生成更自然的攻击样本



Figure. 使用风格化方法生成更自然的样本

对3D物体的实体攻击

交通标识和贴纸仍然是一个平面的对象, 如何对3D实物进行攻击也是一个研究内容. 该攻击成功攻击使得乌龟被误识别识别为步枪.



■ classified as turtle ■ classified as rifle ■ classified as other

Figure. 对抗攻击使乌龟被识别为步枪

对摄像头的实体攻击

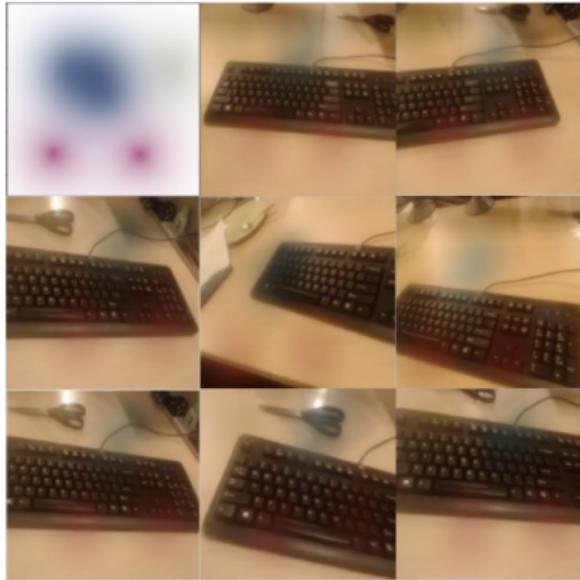


Figure. 在摄像头 上添加贴纸

假如我们并不想攻击哪个具体物体, 只是想让某个摄像头翻车.
但摄像头上的贴纸由于成像原理, 并不会直接成清晰的像, 而是形成大小不一的光斑.

谢谢
